Evaluating the impacts of global environmental assessments

Article (Accepted Version)

Alcamo, J. 2017. Evaluating the impacts of global environmental assessments. *Environmental Science and Policy*. 77(C): 268-272.

Author's accepted manuscript

Final copy-edited version available at:
https://www.sciencedirect.com/science/article/pii/S1462901117302551

Evaluating the impacts of global environmental assessments

Joseph Alcamo

Center for Environmental Systems Research; University of Kassel; Kassel, Germany
Sussex Sustainability Research Programme; University of Sussex; Falmer, UK

Abstract

There are currently no widely accepted procedures for comparing the performance of global environmental assessments (GEAs) and this may be a barrier to improving their methodology. To encourage greater self-reflection within the GEA community, it is proposed to introduce consistent evaluation approaches. Two elements from current evaluation practice are reviewed here that could be particularly useful for evaluating GEAs. The first are logic models which provide a transparent visual mapping of how activities in a GEA are intended to have impacts on policies. The second are performance metrics. It is proposed that GEAs adopt two kinds of metrics: (i) A common generic set for use in all GEAs to provide a basis for comparing the performance of GEAs, and (ii) a specific set of measureable metrics for each particular GEA derived from /linked to the generic set. Although many issues arise in applying these and other elements from evaluation theory and practice to GEAs, the potential benefits are greater comparability of GEA performance and new knowledge about how to improve them. This Short Communication is part of a Special Issue on solution-oriented GEAs.

Keywords

assessment, environmental assessment, evaluation, logic models, performance metrics

1. Introduction

Global environmental assessments (GEAs) are a major tool for synthesizing scientific knowledge of particular relevance to global environmental policymaking. In this way they occupy a vital niche at the interface between global environmental science and policy. A pragmatic indicator of their importance are the large sums that governments regularly invest in them (e.g. the Millennium Ecosystem Assessment cost approximately $25 Million up to 2006; Wells, et al. 2006), and the substantial pro bono time invested by the scientific community in this work. It is argued elsewhere (Kowarsch et al., in review) that GEAs are becoming even more important because of the growing demand for more solution-oriented policy assessments.

For such an important and costly process, it is surprising that the degree of self-reflection within the GEA community is relatively modest. There is only a small literature comparing GEAs (Beck et al. 2014; Leemans 2008; Mitchell et al. 2006a; Rothman et al. 2009) and no agreed-upon procedure or metrics for judging their overall quality. Relatively few evaluations of GEAs[1] have been conducted compared to the number of GEAs, and these have used widely differing approaches making it difficult to compare their results.

One way to raise the level of self-reflection would be for the GEA community to adopt a consistent procedure for impact evaluation. This would allow different assessments to be critically compared and would encourage mutual improvement and development of GEAs. Likewise this would allow the sponsors and stakeholders of recurrent assessments to obtain feedback for improving GEA performance and establishing GEA accountability.

The aim of this Short Communication is to present selected concepts from evaluation theory and practice that can contribute to evaluations of the impacts of GEAs.

2. Programme Evaluation

Ideas for evaluating GEAs can draw on an extensive literature of evaluation theory and practice (e.g. Alkin, 2004; Cardin and Alkin, 2012; Stufflebeam and Shinkfield, 2007; Wholey et al., 2010). An "evaluation" in this literature is "the systematic assessment of the worth or merit of an object." (Stufflebeam and Shinkfield, 2007). Of particular relevance to GEAs are "programme evaluations" (as compared, e.g., to personnel or product evaluations) which are "the application of systematic methods to address questions about program operations and results" (Wholey, 2010). A "programme" in the sense of the evaluation literature is "...a set of resources and activities directed toward one or more common goals, typically under the direction of a single manager or management team" (Wholey, 2010). Defined in this way, GEAs can be seen as a kind of programme, albeit a new kind of programme, in that it is a scientific activity that summarizes rather than develops new knowledge (as in a research programme) with the explicit aim of delivering this knowledge directly to the policy and stakeholder community.

[1] Examples of GEA evaluations: Anon (2007); UNEP/IUCN (2007), Wells et al. (2006).

Although there are several alternative approaches to programme evaluation (as reviewed in Hughes and Nieuwenhuis, 2005; Stufflebeam and Shinkfield, 2007; Wholey, et al. 2010) two elements are common to many different approaches – logic models and performance metrics. Here we focus on these two elements because of their potential value to the task of evaluating GEAs, even if a full evaluation is not performed.

3. Logic models

One of the principal aims of programme evaluation is to judge their impact on intended audiences (Bryson and Patton, 2010). "Impact" in this sense means the positive (and negative) effects produced by a programme; both primary and secondary, direct or indirect, intended or unintended (definition adapted from DAC/OECD, 2002). A common tool used in evaluation practice for mapping impact is the "logic model" (sometimes called a logical framework or "logframe"), defined as a "plausible and sensible model of how a program will work under certain environmental conditions to solve identified problems" (Bickman, 1987). In plain terms, a logic model is a diagram that aims to transparently link programme activities with their impact. A logic model was used in the evaluation of GEO-4 (UNEP/IUCN, 2007), but they are still relatively rare in GEA evaluations.

Logic models come in different forms, but usually include at least a depiction of the activities of a programme, the outputs and outcomes produced by these activities, and the short-, medium- and long-term impacts of these outputs and outcomes. In some cases outputs or outcomes directly generate impacts (e.g. when an assessment report is quoted by a government delegation); in other cases there may be a linear chain of impacts in which short-term impacts generate medium-term impacts, and so on (e.g. when assessment results are first discussed within a government and eventually lead to policy changes in the government). A proposed generic logic model for GEAs is shown in Figure 1. Note this model includes all the components mentioned above, plus an additional one called "assessment processes". It is argued below that not only outcomes but also processes have an impact on the target audiences of assessments.

A general drawback of logic models is that they tend to be superficial (Stufflebeam and Shinkfield), despite the fact that they should be underlain by a "program theory" of how impacts occur (Chen and Rossi, 1983). In practice, developers of logic models more often than not lack the knowledge or theoretical construct to credibly specify the cause and effect of impacts (Newcomer, et al., 2010). This is understandable given the complexity of the impact process and its dependence upon the problem setting (Corbyn, 2011; Weichselgartner and Kasperson, 2010).

Given this drawback, why bother with logic models? The first reason is that they provide a template, albeit imperfect, for making intended assessment impacts visually explicit. Secondly, through this visualization, it becomes easier for evaluators to understand the intended impacts and to discuss them with assessment scientists. Thirdly, developing a logic model forces planners or evaluators of GEAs to make explicit the kind of impacts expected of the assessment. And finally, because of the first three reasons, logic models facilitate comparisons of impacts between assessments, and enable the learning that should come from these comparisons.

Of course, logic models could also be developed in the planning phase of a project (as was done in the Millennium Ecosystem Assessment and Emission Gap reports) to help guide project activities. A logic model developed in the planning phase of an assessment can later be taken over and used as part of an *ex post facto* evaluation.

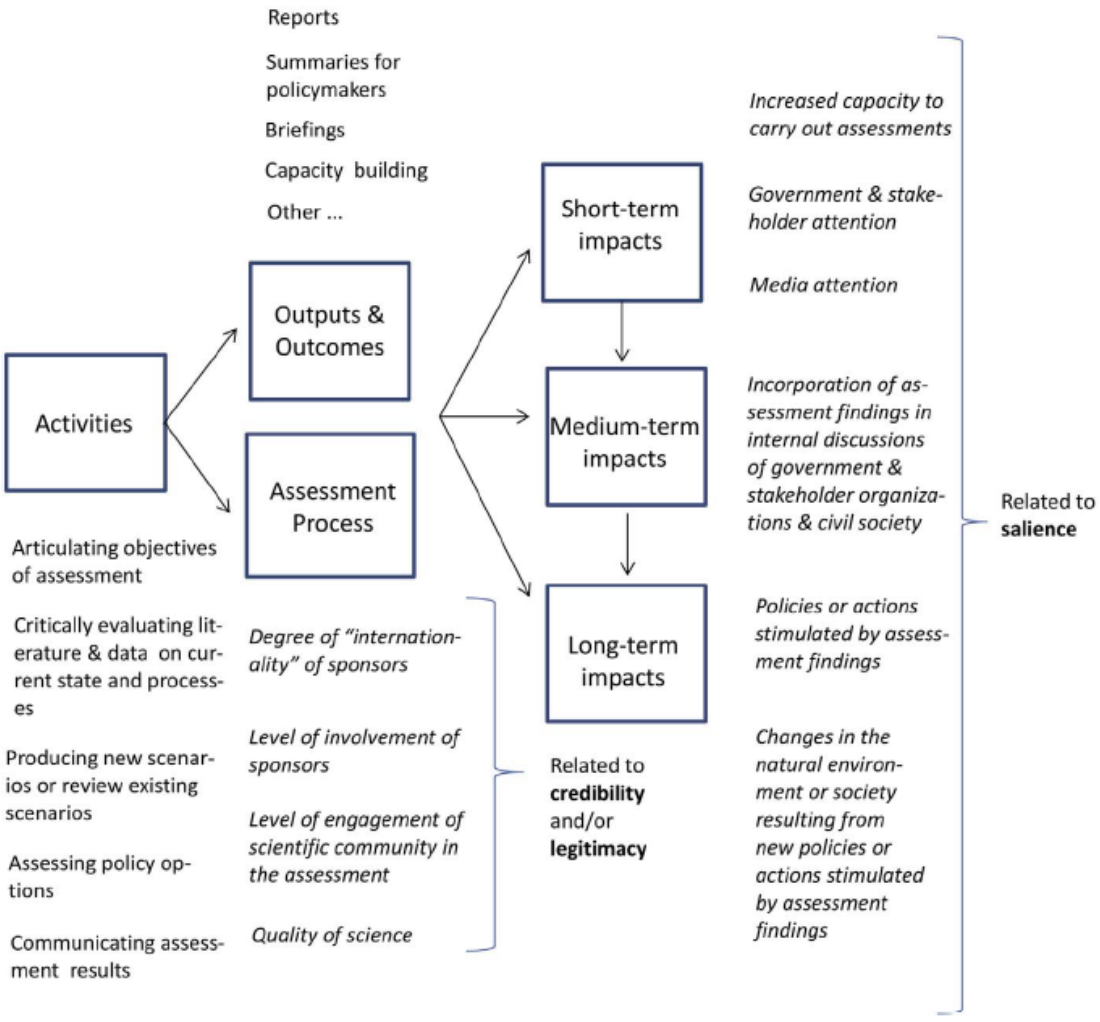

Fig. 1. A suggested generic logic model for global environmental assessments, with generic performance metrics in italics.

4. Performance Metrics

The second element of programme evaluations reviewed here are "performance metrics," defined as "important outcome[s] characteristic[s], attribute[s] or variable[s] of the processes being evaluated" (Hughes and Nieuwenhuis, 2005). Performance metrics (sometimes called "indicators" or "criteria") provide a transparent means of tracking and measuring the impact of a programme (Poister, 2010).

A first step in deriving performance metrics is to identify the characteristics of an assessment that lead to its success. For GEAs, the three characteristics proposed by Clark et al. (2006) have received wide acceptance by the scientific community (e.g. US NRC, 2007; Deri et al. 2009; Rothman et al. 2009). They are: "salience" (or "relevance"), relating to an assessment's ability to communicate with the users whose decisions it seeks to inform and whether the information is perceived as relevant; "credibility" which addresses the technical quality of information, as perceived by the relevant scientific or other expert communities; and "legitimacy" concerning the fairness and impartiality of an assessment process, as judged by its users and stakeholders (definitions from NRC, 2007).

There is an implicit hierarchy here. Since the main motivation for performing global environmental assessments is to provide scientific information and knowledge on the global scale relevant to environmental policy (Rothman et al. 2009), then it is logical that "salience" should be the main basis for judging a GEA's impact (Figure 1). The other two attributes, "legitimacy" and "credibility" are best viewed as prerequisites to achieve salience but not as a mark of success on their own. It follows that a larger number of metrics should be salience-related as compared to the other two characteristics. Note also, that there is difference between how salience and the other two attributes of success are generated within a GEA; whereas salience is achieved *by the outputs and outcomes* of the GEA (e.g. reports and summaries), credibility and legitimacy are generated by the *process* of the assessment (e.g. the quality of science, and the level of participation).

In practice, it is difficult to judge the relative importance of outputs, outcomes or processes in contributing to the final value of an assessment, and it is therefore important to maintain performance metrics for all of these.

Besides the attributes of success, authors in the evaluation literature have proposed other desirable characteristics of metrics. Poister (2010) draws on experience in programme evaluation and suggests the following: (1) validity (does the indicator reflect what is meant to be measured?); (2) reliability (how consistently can the indicator be measured?); meaningfulness and understandability (is the indicator meaningful and understandable to programme managers and/or beneficiaries of the programme?); (3) balance and comprehensiveness (does the indicator provide a balanced and comprehensive view of the programme?); (4) timeliness and actionability (can the indicator be provided in a timely fashion and can programme managers or beneficiaries respond to indicator values?); (5) "degree of goal displacement" (does the indicator lead to overall improved performance or does it lead only to attempts to improve the indicator?); (6) costliness and other practical

considerations including measurability (can the indicator be measured with reasonable effort and costs?). Similar attributes are suggested by Hughes (2005).

In practice, it may be difficult to judge in advance the "reliability and meaningfulness" or "degree of goal displacement" of particular performance metrics; this suggests that a certain period of experimentation will be needed to determine which metrics best fit to these characteristics. Another of the characteristics, "timeliness" is certainly more important for short term rather than medium or long term impacts. That leaves "validity", "balance and comprehensiveness" and "costliness and other practical considerations (measurability)" as particularly important characteristics to consider in the initial selection of metrics.

Table 1 presents two illustrative sets of performance metrics which implicitly consider the above six characteristics, and explicitly consider the attributes of salience, legitimacy and credibility. The first is a generic set which, in principle, should be applicable to all GEAs and serve as a common basis for comparing and tracking the performance of different GEAs. The logic model (Figure 1) discussed earlier provides a useful framework for organizing these metrics. Because of their generality, it is unlikely that the generic metrics will be measureable. This problem is solved if each GEA derives its own set of specific and measureable metrics based on/linked to the generic set (see the example in Table 1). For example, the specific metric "mentions of assessment results in opening speeches at main climate negotiation meetings" could be derived from, and linked to, the generic metric "government & stakeholder attention". Following this approach, each GEA would derive its own specific metrics based on the generic set used by the entire GEA community. Each GEA would then translate the results of its specific metrics to the generic set, thereby making it possible to compare the performance of different GEAs. Exactly how to do this translation is an open question, and may require, for example, the setting of thresholds on metric values (see below). Another important issue is the acceptance and legitimacy of these metrics. One option for handling all of these issues is to develop the generic metrics through a participative process involving both the GEA community-at-large and stakeholders.

Other evaluation issues

The preceding paragraphs raise only a few of the many issues involved in applying logic models, performance metrics, and other elements of evaluation theory and practice to GEAs. Also important are institutional aspects of GEA evaluations, including the utilization of evaluation results. Experience shows that even well-designed evaluations are sometimes/often underutilized. One reason could be the lack of ownership on the part of evaluation stakeholders (World Bank, 2009). This could stem from the fact that GEA evaluations, if they are carried out at all, are often poorly integrated with the rest of the planning of a GEA. Not infrequently, their essential role in the

Table 1 Example performance metrics for GEAs. The first column indicates the particular GEA attribute ("salience, legitimacy, credibility") which the metric addresses. The second column shows generic metrics that may be applicable for all GEAs. The time scale of impact in parentheses refers to time scales in Figure 1. For illustration, the third column shows suggestions for specific metrics that follow from the generic ones. As an example, suggestions for performance metrics for UNEP Emission Gap reports (UNEP, 2010) are shown.

GEA attribute	**Possible generic performance metrics for GEAs**	**Suggestions for specific performance metrics (related to UNEP Emission Gap reports)**
Salience	Increased capacity to carry out assessments	Number of young scientists from national climate offices trained to estimate national emissions.
Salience	Government & stakeholder attention (short term impact)**	Mentions of assessment results in opening speeches at main climate negotiation meetings * Number of downloads of assessment report (if possible, by government and stakeholder organizations).
Salience	Media attention (short term impact)	Number of mentions of report findings by major media within (5) days following launch of assessment report* Number of interviews by major media regarding assessment results
Salience	Degree to which assessment findings are incorporated into internal discussions of government & stakeholder organizations (medium term impact)	Number of government departments in different countries using the assessment report. Number of invitations to present assessment findings at workshops/events sponsored by governments or other stakeholders.
Salience	Number of policies and actions that are stimulated by assessment findings (short to long term impact)	Number of government or stakeholder policies or policy statements that refer to assessment findings.
Salience	Changes in the natural environment or society resulting from new policies or actions stimulated by assessment findings (long term impact)	Reductions in greenhouse gas emissions that can be traced back to policies or policy statements referring to assessment findings
Legitimacy	Degree of "internationality" of sponsors or sponsoring organizations (short term impact)	Number of countries represented in sponsoring organization of the assessment (United Nations Environmental Assembly) Number of countries represented in steering committee of report.
Legitimacy	Level of involvement of sponsors or other potential users (short term impact)	Number of meetings between scientists doing assessment and sponsors of assessment. Number of countries or stakeholder organizations submitting comments on assessment reports.
Credibility	Level of engagement of scientific community in the assessment (short term impact)	Total number of research groups active in assessment. (analysis, writing, review).* Number of research groups from different countries active in assessment (analysis, writing, review).*
Credibility	Quality of science (short to long term impact)	Number of reviewers in scientific peer-review process of assessment Number of citations in peer-reviewed journals of assessment results.

* Items marked with an asterisk were used in some capacity by the author and his institution in the informal evaluation of the Emission Gap reports (UNEP, 2010). Suggestions without an asterisk appear for the first time in this Short Communication.

** Short-, medium-, and long-term impacts refer to different time scale of impacts depicted in Figure 1. In this Short Communication, "short-term" refers to the period up to the publishing of the main assessment results, "long-term" is the period at least three years after the main assessment results are published, and "medium-term" is the period in-between.

ongoing improvement of performance and accountability of the GEA is overlooked (or not accepted) by sponsoring institutions and other stakeholders. One way, then, to strengthen the sense of ownership would be to institutionalize evaluations – i.e. make them a part of a regular ongoing GEA process, endow them with as much importance as other aspects of the process, integrate them in the work programmes of the sponsoring organizations, and provide adequate resources to keep stakeholders fully informed about their findings.

Other important questions are:

- Will evaluation, including the development of logic models, become so complicated that disproportional resources go into the evaluation rather than into what is being evaluated? (McLaughlin and Jordan, 2010).
- Is it appropriate to apply a uniform set of performance metrics to a heterogeneous set of GEAs, made up of varying goals, scopes, methods, and other characteristics? The proposed two-level approach to metrics may address this question. The generic metrics are general enough to be applied to all/most GEAs, while the specific-metrics take into account their heterogeneity. Nevertheless, the effectiveness of this approach has yet to be tested in practice.
- What is the best way to acquire data about performance metrics? There has been a lively debate in the evaluation community as to whether "scientific" survey methods are necessary or feasible for acquiring data about metrics (Stufflebeam and Shinkfield, 2007).
- Thresholds could be helpful for establishing whether a specific metric level is "good", "very good", and so on. This valuation would, in turn, be used to translate a specific metric level to a generic metric level, and would make it possible to compare the performance of different GEAs. But how should thresholds for performance metrics be set?

While it may be difficult to address these and other issues, the potential benefits from resolving them and introducing a consistent evaluation approach to GEAs are great – An enhanced ability to compare the performances of GEAs, and greater understanding of how to make this already useful methodology even more beneficial to environmental policy and the needs of society.

Acknowledgements. The author is indebted to B. Alcamo, M. Kowarsch, J. Jabbour, and an anonymous reviewer for their helpful comments on an earlier draft of this Communication.

References

Alkin, M. (Ed.), 2004. Evaluation Roots: Tracing Theorists' Views and Influences. Sage, Thousand Oaks, CA.

Anon, 2007. Implications of findings of the Millennium Ecosystem Assessment on the work of the Convention. Subsidiary Body on Scientific, Technical and Technological Advice of the Biodiversity Convention. UNEP/CBD/SBSTTA/12/4.

Beck, S. et al., 2014. Towards a reflexive turn in the governance of global environmental expertise. The cases of the IPCC and the IPBES. GAIA Journal. 23(2), 80-87

Bickman, 1987, quoted in: McLaughlin, J., Jordan, G., 2010. Using logic models, in: Wholey, J., Hatry, H., Newcomer, K. (eds.) Handbook of Practical Program Evaluation. 3rd Edition. Jossey-Bass. San Francisco. pp. 55-80

Bryson, J., Patton, M., 2010, Analyzing and engaging stakeholders, in: Wholey, J., Hatry, H., Newcomer, K. (eds.) Handbook of Practical Program Evaluation. 3rd Edition. Jossey-Bass. San Francisco. pp. 30-54.

Cardin, F. Alkin, M., 2012. Evaluation roots: an international perspective. Journal of Multi-Disciplinary Evaluation. 8(17), 102-118.

Chen, H., & Rossi, P., 1983. Evaluating with sense: the theory-driven approach. Evaluation Review. 7, 283-302.

Clark, W., Mitchell, R., Cash, D. (Eds.), 2006. Evaluating the influence of global environmental assessments, in: Mitchell, R., Clark, W., Cash, D. (Eds.) Global Environmental Assessments: Information and Influence. MIT press, Cambridge, Massachusetts. 1-28.

DAC/OECD, 2002. Quality standards for development evaluation. DAC Guidelines and Reference Series. OECD, Paris.

Déri, A., Swanson, D., Bhandari, P., 2009. IEA (Integrated Environmental Assessment) Training Manual - Monitoring, Evaluation and Learning. United Nations Environmental Programme. Nairobi, Kenya.

Donaldson, S., Lipsey, M. Roles for theory in evaluation practice. In: Shaw, I., Greene, J., Mark, M. (eds.) 2006. The SAGE Handbook of Evaluation. SAGE, Thousand Oaks.

Hughes, J., Nieuwenhuis, L., 2005. A Project Manager's Guide to Evaluation. Evaluate Europe Handbook Series Volume 1. European Commission.
http://www.pontydysgu.org/wp-content/uploads/2008/02/EvaluateEuropeVolume1final.pdf

Kowarsch, M., Jabbour, J., Flachsland, C., Kok, M., Watson, R., Haas, P. Minx, J., Alcamo, J. Garard, J., Riousset, P., Pintér, L., Langford, C., Yamineva, Y., von Stechow, C., O'Reilly, J., Edenhofer, O. Global environmental assessments and the path to solutions. In review *Nature Climate Change.*

Leemans, Rik, 2008. Personal experiences with the governance of the policy-relevant IPCC and Millennium Ecosystem Assessments. Global Environmental Change. 18, 12–17.

Leeuw, F., Donaldson, S., 2015. Theory in evaluation: Reducing confusion and encouraging debate. Evaluation. 21(4), 467–480

McLaughlin, J., Jordan, G., 2010. Using logic models, in: Wholey, J., Hatry, H., Newcomer, K. (eds.) Handbook of Practical Program Evaluation. 3rd Edition. Jossey-Bass. San Francisco. 55-80

Mitchell, R., Clark, W., Cash, D. (Eds.) 2006a. Global environmental assessments: information and influence. MIT press, Cambridge, Massachusetts.

Mitchell, R., Clark, W., Cash, D., 2006b. Information and influence, in: Mitchell, R., Clark, W., Cash, D. (Eds.), Global environmental assessments: information and influence. MIT press, Cambridge, Massachusetts. 322-361.

Newcomer, K., Hatry, H.. Wholey, J., 2010, Planning and designing useful evaluations, in: Wholey, J., Hatry, H., Newcomer, K. (Eds.) Handbook of Practical Program Evaluation. 3rd Edition. Jossey-Bass. San Francisco. 5-29

Poister, T. 2010. Performance measurement, in: Wholey, J., Hatry, H., Newcomer, K. (Eds.) Handbook of Practical Program Evaluation. 3rd Edition. Jossey-Bass. San Francisco. 100-124

Rothman, D., van Bers, C., Bakkes, J. Pahl-Wostl, C., 2009. How to make global assessments more effective: lessons from the assessment community. Current Opinion in Environmental Sustainability.1, 214–218

Stufflebeam, D., Shinkfield, A. 2007. Evaluation theory, models, and applications. Jossey-Bass, San Francisco, USA.

Stufflebeam, D., 1983. The CIPP model for program evaluation. In: Madaus, G. Scriven, M., Stufflebeam, D. (eds.), Evaluation models: Viewpoints on Educational and Human Services Evaluation Boston: Kluwer-Nijhoff. 117-141.

Suchman, E., 1967. Evaluative Research: Principles and Ppractice in Public Service and Social Action Programs. Russel Sage, New York.

UNEP/IUCN, 2007. Review of the initial impact of the GEO-4 report. United Nations Environmental Programme. Nairobi, Kenya.

UNEP, 2010. The Emissions Gap Report. United Nations Environmental Programme. Nairobi, Kenya.

US NRC (National Research Council), 2007. Analysis of Global Change Assessments: Lessons Learned. Committee on Analysis of Global Change Assessments, www.nap.edu.

Weichselgartner, J., Kasperson, R., 2010. Barriers in the science-policy-practice interface: Toward a knowledge-action-system in global environmental change research. Global Environmental Change. 20, 266–277

Wells, M. Grossman, D., Navajas, H., 2006. The terminal evaluation of the "Millennium Ecosystem Assessment". Prepared by UNEP Evaluation and Oversight Unit. United Nations Environmental Programme. Nairobi, Kenya.

Wholey, J., Hatry, H., Newcomer, K. (eds.) 2010. Handbook of Practical Program Evaluation. 3rd Edition. Jossey-Bass. San Francisco.

World Bank. 2009. Institutionalizing impact evaluation within the framework of a monitoring and evaluation system. Independent Evaluation Group. The World Bank Group, Washington, D.C.